◆特集 (M&BE, Vol.30, No.4 (2019)) (原稿受理日：2019 年 10 月 25 日)
「有機分子・バイオエレクトロニクスとマテリアルインフォマティクス」

グラフ畳み込みニューラルネットワークによる立体配座—エネルギー相関

立花尚登，井上雅都，金子晶夫，加藤凱生，濱田信次，○後藤仁志
豊橋技術科学大学 大学院工学研究科 情報・知能工学系
gotoh@tut.jp

1. はじめに

定量的構造物性相関（QSPR）は化合物の構造と物性の関係をケモインフォマティクスの技術を使ってモデル化する手法である。その中でも深層ニューラルネットワーク（DNN: Deep Neural Networks）を用いた QSPR は，理論計算と比較して計算コストが低く，高速かつ十分な精度で膨大な数の化合物の物性評価を可能にすることが期待されている。最近では，二次元分子構造を無向グラフとして畳み込み演算を行うグラフ畳み込みニューラルネット（GCNN: Graph Convolutional Neural Net）にも注目が集まっている。しかし，物質材料の物性は三次元原子配置に依存する相互作用や電子構造によって決まるため，より高い精度で物性予測を行うためには立体構造の情報をどのように機械学習に取り込むかが大きな課題となっている。特に，有機分子の物性は複数の配座異性体を考慮することも多く，それらの立体構造も重要な情報として取り込む必要がある。

我々は，主に定量的活性相関（QSAR）の予測精度の向上を目指して[1]，配座データベース C3DB（Computational Chemistry Conformation Database）の開発を続けている[2]。本稿では，この C3DB を QSPR に応用した研究事例[3]の一部を紹介する。C3DB は，比較的小さな有機分子の SMILES と立体構造の情報を公開データベースから取得し，独自に開発した CONFLEX[4]を用いて配座異性体を創出し，量子化学計算を適用した結果等を収録している。ここでは, C3DB から抽出した立体構造情報と電子構造計算による全エネルギー, HOMO, LUMO, HOMO-LUMO Gap に関する情報を用い, 三次元記述子を入力とした DNN モデルと 3 種類の GCNN モデルによる予測結果を比較する。また，配座異性体を考慮した複合モデルによる全エネルギー予測の結果についても紹介する。

2. 方法

2.1 配座データベース C3DB

C3DB は，米国 NIH の化合物データベース PubChem[5]，および理化学研究所で開発している量子化学データベース PubChemQC[6]から分子情報を取得し，汎用分子計算プログラム CONFLEX[4]を用いて分子力学計算による立体構造，および配座異性体を創出し，さらに，その中から安定な配座異性体に対して量子化学計算プログラム GAMESS[7]や Gaussian[8]などを用いて構造最適化を行い，そこまでに得られた全ての立体構造情報（座標や各種エネルギーなど）を自動的に収録するシステムである。C3DB システムの特徴は，立体配座を徹底的に探索することによって最安定配座だけでなく，熱力学的に安定なすべての配座異性体を収録していることである。

2.2 分子データセット

本研究では，C3DB に収録されている構成元素に H, C, N, O, F, P, S, Cl, Br, I が含まれ，分子量 100 以下の 6,955 分子を抽出し，それの最安定配座の立体構造情報と物性値を最

安定データセットとする。また，最安定データセット内の分子の配座異性体をエネルギーの低い順に最大 10 配座まで抽出したところ，計 25,118 配座の立体構造情報と物性値が得られ，これを配座データセットとする。ここで物性値には，Gaussian [8]を用いて求めた B3LYP/6-31G*レベルの全エネルギー，HOMO，LUMO，HOMO-LUMO Gap を採用する。

機械学習のためにデータセット内のデータを訓練用（Train セット）と評価用（Test セット）に分割する。まず，最安定データセットを Train set と Test set の比が 4:1 になるように分割する。次に，配座データセットに対して，最安定データセットの Train セットと Test セットに含まれる分子の配座異性体を Train セットと Test セットに分類する。ここでは，最安定データセットの分割をランダムに 3 回行い，それに従って配座データセットを分割することで，それぞれ 3 種類のデータセットを生成する。これらをデータセット 1，データセット 2，データセット 3 と呼ぶことにする。

2.3　三次元分子記述と深層ニューラルネット

最安定データセットの Train セット内の分子を 3 次元記述子である 3D-MoRSE [9]で変換した Fingerprint vector を入力にして，Single-task DNN モデルを用いて全エネルギー，HOMO，LUMO，HOMO-LUMO Gap を目標値として学習し，Test セット内の分子の予測値を求める。Single-task DNN とは，出力層に一つのニューロンのみを持つ全接続フィードフォワードネットワークのことで，一つの物性値のみを予測する DNN モデルのことである。その概要を図 1 に示した。ここでは，各分子は 3D-MoRSE で変換された固定長 155 次元の Fingerprint vector として入力され，それぞれ 3100，1550，775 個のニューロンで構成される三段階の隠れ層を経て，一つの予測値が出力される。各ニューロンでは，次式によって出力値（output）が評価され，次層の全ニューロンに渡される。

$$\mathrm{output}_i = f\left(\sum_j \mathrm{w}_{\mathrm{ij}} x_j + b_i\right) \qquad (1)$$

ここで，x_jは前層のニューロン j の値，w_{ij}は荷重パラメータ，b_iはバイアスパラメータである。活性化関数 fは様々な関数が適用できるが，ここでは Sigmoid 関数を用いた。訓練セットの全分子の予測値は目標値との誤差が計算され，誤差が小さくなるようwと b を更新することで DNN モデルが最適化される。今回は Adam 法 [10]を適用した。

3D-MoRSE は Schuur と Gasteiger らによって提案された 3 次元分子記述子であり [9]，電子線回折に基づいた分子表現として次式で表される：

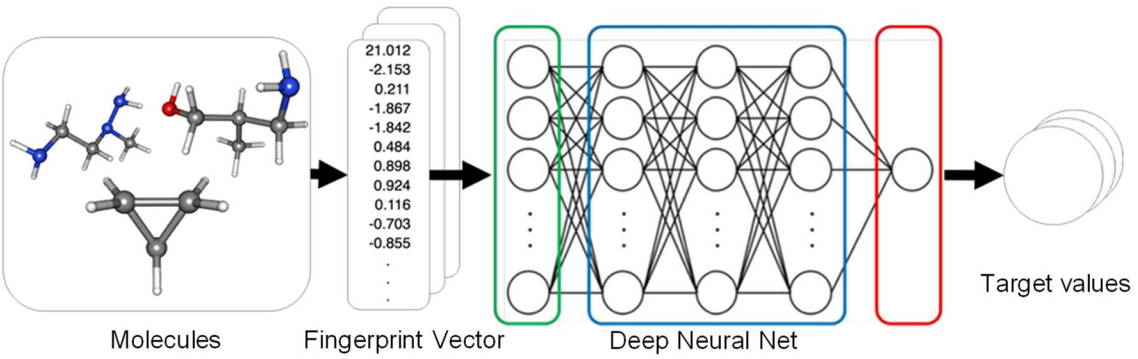

図 1　Single-task DNN モデルの概要図

$$I(s) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} A_i A_j \frac{\sin sr_{ij}}{sr_{ij}},\ s = 0: \lim_{s \to 0} \frac{\sin \quad _{ij}}{sr_{ij}} = 1 \qquad (2)$$

ここで，r_{ij}は原子 i と j の距離，N は原子数，A_iとA_jは重み付けファクターとして用意された原子の特徴量で，非荷重（$A_i = A_j = 1$），原子質量，ファンデルワールス体積，電気陰性度，分極率の 5 種類を用いる。また，s は散乱パラメータであり，0 から 31 の整数値が代入され，32 本のスペクトルが計算される。ただし，s=0 は s→0 の極限値とする。

2.4 グラフ畳み込みニューラルネット

グラフ畳み込みニューラルネットは，グラフの局所構造に対して非線形変換を繰り返すことで，局所構造情報を固定次元の連続値ベクトルに変換する手法である。つまり，通常の DNN が各層のニューロンを全接続しているのに対して，GCNN はニューロンを原子と見立てて，結合している原子間を接続する。また，DNN では隠れ層の数を増やすことで複雑な特徴量を表現しようとするが，GCNN では畳み込みモジュールを再帰的に繰り返すことで，複雑な特徴量を原子に帰属させる。本研究では NFP (Neural Fingerprint) [11]，GGN (Gated Graph Sequence Neural Networks) [12]，MGC (Molecular Graph Convolutions) [13]のを用いた。NFP は，分子内原子の特徴量に周辺の原子の特徴量を畳み込む作業を繰り返して原子の特徴量を更新する手法である [11]。結合は隣接原子を特定するために使われるだけで，結合の特徴量を明示的に考慮しない点で GGN や MGC と異なる。GGN は，最新の Gated Recurrent Unit（GRU）技術を用いている点で特徴的であり，原子情報に加えて結合情報も畳み込みに用いる [12]。また，分子グラフ中の各ノードが有向エッジで接続し，エッジの向きを考慮した隣接行列を用いてノード更新を行う。MGC は，原子の特徴量と結合の特徴量を明示的に定義し，原子の特徴量に結合の特徴量を織り込む Weave Net モジュールが特徴的である [13]。これら GCNN の各手法の詳細は各文献を参照されたい。

尚，DNN と GCNN の学習は，全て汎用 DNN プラットフォーム Chainer [14]，および Chainer Chemistry ライブラリ [15]を用いた。

2.5 精度の評価

構築したモデルの予測精度の評価は，Test セットに対する決定係数 R2 値（式 3）と平均二乗誤差平方根（RMSE）（式 4）を用いる：

$$R^2 = \frac{[\sum_{i=1}^{N}(x_i - \bar{x})\,(y_i - \bar{y})]^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (3) \qquad \mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}} \quad (4)$$

ここで，N はデータ数，x_iとy_iはそれぞれ予測値と目標値（理論値），$\bar{x}$と$\bar{y}$は平均値である。訓練セットを用いた DNN モデルの最適化では，平均二乗誤差（MSE）が使われる。

3. 結果と考察

3.1 最安定データセットの全エネルギー

図 1 は，最安定データセット 1 の Train セットを用いて最適化した(a) 3D-MoRSE/DNN, (b) NFP，(c) GGN，(d) MGC モデルによる Test セットの予測値を理論値と比較した相関グラフである。どの方法でも高い相関を示したことが分かる。3 回ずつ行ったデータセット 1 から 3 の決定係数 R2 の平均パフォーマンスは，それぞれ，0.950，0.988，0.996，0.995 で，GGN と MGC モデルによる予測精度はとても優れていた。しかし，平均 RMSE 値で

は，それぞれ，22.9，11.5，5.63，6.56 hartree であり，化学研究に利用できる精度には，未だ達していない。

3.2　最安定データセットの HOMO，LUMO，HOMO-LUMOGap

軌道エネルギーではどうだろう？図２は，最安定データセット 1 の Train セットを用いた学習後の各モデルの HOMO，LUMO，HOMO-LUMO Gap の予測値と理論値の相関を表している。この図から明らかなように，3D-MoRSE/DNN モデルが高い予測精度を示し，実際，平均 R2 は，順に 0.816，0.873，0.811 であった。これに対して，GCNN 系では，原子特徴量しか考慮しない NFP のパフォーマンスが著しく悪く，いずれも 0.3～0.4 程度であった。一方，結合の特徴量が考慮された GGN の R2 パフォーマンスは，それぞれ 0.741，0.862，0.754 で，同じ様に結合特徴量が織り込まれた MGC のそれよりも若干良くなっており，さらに LUMO エネルギーに関しては 3D-MoRSE/DNN モデルよりも優れていた。

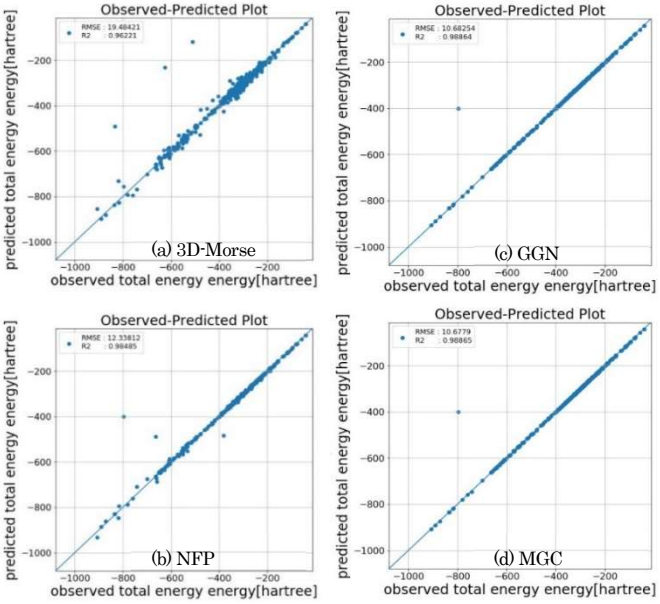

図１　DNN と GCNN モデルによる全エネルギーの予測値と理論値（B3LYP/6-31G*）との比較（最安定データセット 1）：(a) 3D-MoRSE/DNN，(b) NFP/GCNN，(c) GGN/GCNN，(d) MGC/GCNN

平均 RMSE は，3D-MoRSE/DNN で，それぞれ 0.476，0.543，0.754 eV となった。当初，我々は，LUMO エネルギーの予測は難しいだろうと考えていたが，今回の結果では，LUMO エネルギーの予測値と理論値の相関性は，HOMO や Gap よりも高かった。一方，LUMO の RMSE は比較的大きく，もう少し解析を進めてみる予定である。また，Gap の誤差が HOMO や LUMO よりも大きくなったことにも注意する必要があるだろう。この点もさらに解析が必要であるが，10 万件以上の構造データを用いた同様な研究と比較しても，入力データ数の違いを考慮すれば，我々の結果は妥当であると思われる [16]。

3.3　配座データセットの全エネルギー

配座データセット 1 の Train セットを用いた学習後の 3D-MoRSE/DNN の全エネルギー予測値と理論値との比較を図 3(a)に示した。また，複合モデルの各配座の全エネルギーの予測値$E_{EM,composite}$は，その 3D-MoRSE/DNN の予測値から求めた配座エネルギー（$E_{EM,DNN} - E_{GM,DNN}$）に，最安定データセットの GCNN モデルの予測値$E_{GM,GCNN}$に加えた値である（式(5)）。各 GCNN モデルの予測値と理論値の比較を図 3(b)～(d)に示した。

$$E_{EM,composite} = E_{GM,GCNN} + \left(E_{EM,DNN} - E_{GM,DNN}\right) \quad (5)$$

それぞれの平均 R2 パフォーマンスは，0.973，0.995，0.989，0.995 となった。3D-MoRSE/DNN よりも各 GCNN と複合した方が良い予測精度だったことは，前述の結果から理解できる。ところが，最安定データセットを用いた際の予測精度が他と比べて良くなかった NFP モデルの結果と複合させた NFP+3D-MoRSE の予測値が，比較的高いパフォーマンスを示していたことはとても興味深い。現在，さらなる解析を進めている。

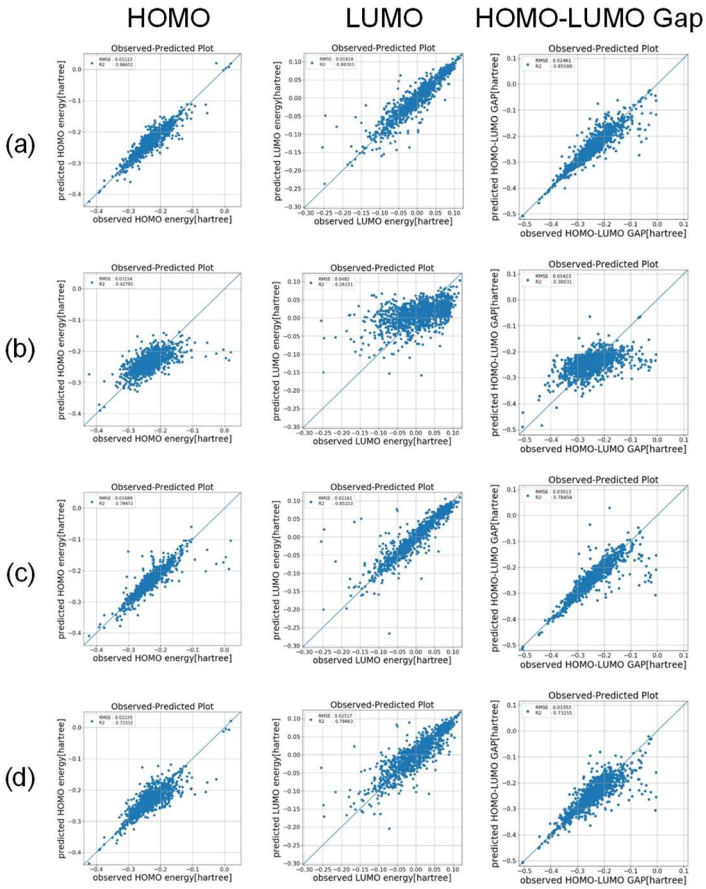

図２　DNN と GCNN モデルによる HOMO, LUMO, HOMO-LUMO Gap の予測値と理論値（B3LYP/6-31G*）との比較（最安定データセット 1）: (a) 3D-MoRSE/DNN, (b) NFP/GCNN, (c) GGN/GCNN,　(d) MGC/GCNN

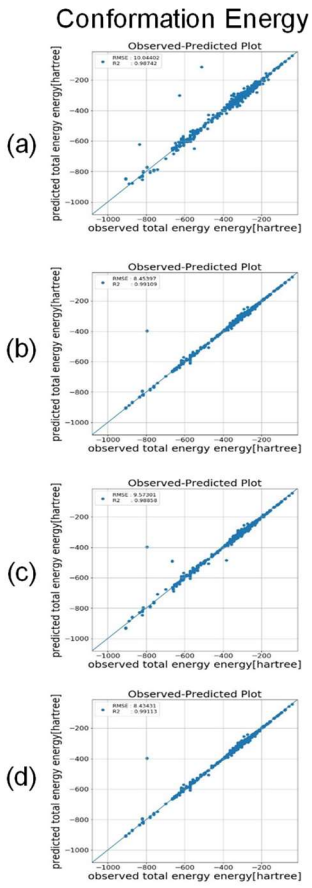

図 3　全エネルギーの予測値と理論値(B3LYP/6-31G*)の比較（配座データセット 1）: (a) 3D-MoRSE, (b) NFP+3D-MoRSE, (c) GGN+3D-MoRSE, (d) MGC+3D-MoRSE

4. まとめと今後の展望

本稿では，我々が QSAR 研究のために開発してきた C3DB を QSPR へ応用した研究事例を紹介した。ここでは，理論計算による物性評価の基礎である全エネルギーや軌道エネルギーについて，6,955 分子の最安定構造と計 25,118 配座異性体の立体構造情報を用いて，DNN と GCNN モデルを用いて予測値を求めた。その結果，B3LYP/6-31G* レベルの理論値との比較では，最安定構造の全エネルギーに対する R2 パフォーマンスはどのモデルでも良い相関（平均 R2＞0.95）を示した。しかし，化学研究に活用するには，RMSE をさらに改善する必要があると思われる。DNN と GCNN モデルを組み合わせた配座異性体の全エネルギーの予測は，最安定構造のそれよりも良い相関を示したことは興味深い。

以上の結果から，全エネルギーについては，GCNN のように分子構造をグラフ表現したモデル，つまり二次元構造情報だけでも十分な予測精度を導き出すことが期待できると思われる。一方，軌道エネルギーの予測には，三次元構造情報が重要であることが明白である。特に，GCNN の結果からは，原子の特徴量だけでは明らかに不十分ではあるが，化学結合の特徴量を丁寧に畳み込んだ新たな特徴量を導出することができれば，さらなる精度改善が見込まれる。ただし，残念なことに，DNN はもとより GCNN においてはさ

らに，ネットワーク内に暗に構築されているであろう有効な特徴量を，我々が理解できる形式にすることができない。現在，この点に注目した新たな方法論の開発が進んでおり，今後の展開として期待しているところである。

本稿で紹介した結果と解析は，中間報告としてまとめたものであることに留意されたい。現在，さらにデータセットを増やし，初期パラメータを変えた各モデルの最適化を複数回実行した結果を解析中であり，ここでは不明だったいくつかの点が明らかになりつつある。詳細については，速やかに発表する予定である [2)]。

本研究の一部は日本学術振興会科学研究費補助金（JP17H06373）より支援されました。Y. K.は文部科学省リーディング大学院プログラム「超大規模脳情報を高度に技術するブレイン情報アーキテクトの育成プログラム」を通じて支援されています。また，機械学習の一部は、東京工業大学の TSUBAME3.0 スーパーコンピューターで実行されました。

参考文献

1) Y. Kato, S. Hamada, H. Goto, Proceedings of International Conference on Advanced Informatics: Concepts, Theory And Application (ICAICTA), 2016, 1-6.
2) N. Tachibana, M. Inoue, A. Kaneko, Y. Kato, S. Hamada, H. Goto, Manuscript in preparation.
3) Y. Kato, S. Hamada, H. Goto, Manuscript submitted for publication.
4) H. Gotō, E. Ōsawa,, *J. Am. Chem. Soc.* 111 (1989), 8950-8951. H. Gotō, E. Ōsawa,, *J. Chem. Soc., Perkin Transaction*s 2 (1993), 187.
5) S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, E. E. Bolton, *Nucleic Acids Research* 47 (2018), 47, 1102-1109.
6) M. Nakata, T. Shimazaki, *J. Chem. Inf. Model.* 57 (2017), 1300-1308.
7) M. S. Gordon, M. W. Schmidt, "Theory and Applications of Computational Chemistry: the first forty years", Elsevier, Amsterdam, 2005, pp.1167-1189,
8) Gaussian 16, Revision A.03, M. J. Frisch, *et. al*., Gaussian, Inc., Wallingford CT, 2016.
9) O. Devinyak, D. Havrylyuk, R. Lesyk, J. Mol. Graph. Model., 54 (2014), 194-203.
10) D. P. B. Kingma, Jimmy Lei, the 3rd International Conference for Learning Representations, arXiv preprint (2015), arXiv:1412.6980v8.
11) D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, R. P. Adams, *arXiv preprint* (2015), arXiv:1509.09292v2.
12) Li, Y., Tarlow, D., Brockschmidt, M., & Zemel, R., *arXiv preprint* (2015), arXiv:1511.05493.
13) S. Kearnes, K. McCloskey, M. Berndl, V. Pande, P. Riley, *J. Comput. Aided Mol. Des.* , 30 (2016), 595. *arXiv preprint* (2016), arXiv:1603.00856.
14) S. Tokui, K. Oono, S. Hido, J. Clayton, Proceedings of workshop on machine learning systems in the 29th annual conference on neural information processing systems (NIPS), 5 (2015), 1-6.
15) S. Tokui, R. Okuta, T. Akiba, Y. Niitani, T. Ogawa, S. Saito, S. Suzuki, K. Uenishi, B. Vogel, H. Y. Vincent, KDD '19 Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2019), 2002-2011.
16) F. Pereira, K. Xiao, D. A. R. S. Latino, C. Wu, Q. Zhang, J. Aires-de-Sousa, J. Chem. Inf. Model. 57 (2017), 11-21.